Supporting Information. Kass, J.M., S.I. Meenan, N. Tinoco, S.F. Burneo, and R.P. Anderson. 2020. Improving area of occupancy estimates for parapatric species using distribution models and support vector machines. Ecological Applications.

Appendix S2: Species distribution modeling details

We chose a subset of the WorldClim 2.0 set of 19 bioclimatic variables for use as predictor variables in our SDMs: mean diurnal range (bio02), temperature seasonality (bio04), minimum temperature of coldest month (bio06), mean precipitation of wettest month (bio13), mean precipitation of driest month (bio14), precipitation seasonality (bio15), mean precipitation of warmest quarter (bio18), and mean precipitation of coldest quarter (bio19). To avoid collinearity in this subset of predictor variables and aid interpretability of the results, we removed those with a high variance inflation factor (VIF) in a step-wise fashion until none exceeded a VIF of 10, using the R package usdm (Naimi *et al.*, 2014)). This process was done after masking all variable rasters by a shared extent, defined by a minimum convex polygon around the occurrence localities for both species, buffered by 50 km.

As both species' occurrence data had (presumably artifactual) spatial clustering that could bias model results (Veloz, 2009), we thinned occurrence localities for each species separately by 5 km (to both minimize data loss and remove obvious clusters) using functionality from the R package spThin (Aiello-Lammens *et al.*, 2015). We downloaded the bioclimatic variables we selected above at 30 arcsec resolution (1 km at the equator) in *Wallace* 1.0.6 (Kass *et al.*, 2018). We delineated study extents for each species defined by a convex hull (i.e., minimum convex polygon) around the occurrence localities buffered by 50 km, a distance that captures as much of the area likely available to each species without including regions east of the crest of the Andes, where neither species has been found (and to which presumably neither were historically able to disperse). We created these study extent polygons in R because *Wallace* cannot currently buffer extents in meters, and input these shapefiles as a user-specified study extent to mask the bioclimatic rasters. We extracted background values for every cell with climatic data from each species' respective study extent in order to get a comprehensive background sample to avoid artifactually truncating the model response (Guevara *et al.*, 2018).

We built SDMs within *Wallace* using Maxent 3.4.1, a presence/background, machine learning modeling method that estimates a species' response to environmental predictor variables subject to constraints derived from these variables (Phillips *et al.*, 2017). Maxent models can potentially fit very complex responses, but complexity can be increasingly penalized to result in simpler responses—these settings can be tuned to allow for the selection of models with high performance and low overfitting (Merow *et al.*, 2013; Radosavljevic & Anderson, 2014). In Maxent, feature classes control the various shapes of the modeled response and determine how complex it can be, and higher values of the regularization multiplier enforce simpler models with fewer resulting non-zero coefficients (Phillips & Dudík, 2008; Merow *et al.*, 2013). For each species, we built suites of models with varying levels of complexity based on functionality provided from the R package ENMeval 0.3.0 (Muscarella *et al.*, 2014), considering combinations of linear (L), quadratic (Q), and hinge (H; similar to splines) feature classes (L, LQ, H, and LQH) with a range of regularization multipliers (0.5 to 5 by increments of 0.5). This resulted in 40 candidate models per species. For evaluation, we implemented the $n-1$ "jackknife" method (or "leave-one-out"), which is recommended for maximizing the

information available for model training when the sample size is small (Pearson *et al.*, 2007; Shcheglovitova & Anderson, 2013).

We selected optimal models sequentially by choosing those that accurately predicted the most withheld occurrence localities (i.e., lowest average omission rate), and then to break any ties, those with the best discriminatory ability on the withheld occurrences (highest average test AUC). We first removed from consideration all models with fewer than two parameters (non-zero coefficients), as these had extremely unrealistic predictions and responses, and those with average AUC_{test} below 0.5, which indicates poor discriminatory ability (see description below). We then selected models with the lowest average omission rate on withheld data. As the occurrence data for both species was derived from taxonomic identifications made by specialists and nearly all had relatively low estimated spatial error (<5 km), we chose the minimum training presence (MTP) threshold, which is based on the minimum suitability value of the training data. When multiple candidate models were tied for lowest omission rate, we selected the one with the highest average AUC_{test}. The area under the curve (AUC) of the receiver operating characteristic is a standard measure of discriminatory ability for SDMs (Fielding & Bell, 1997), providing a threshold-independent evaluation of the model's ability to differentiate presences from absences (or in this case, background; Peterson *et al.*, 2011), and AUC_{test} is the AUC calculated on withheld test data. Although there are problems with interpreting AUC in absolute terms as a measure of accuracy for presence/background models (Lobo *et al.*, 2008), it is a valid metric to compare among models for a single species across the same study extent (Peterson *et al.*, 2011). We calculated both omission rate and AUC_{test} on each withheld record in turn, and took the average across all iterations (Shcheglovitova & Anderson, 2013).

Literature Cited

Aiello-Lammens, M.E., Boria, R.A., Radosavljevic, A., Vilela, B. & Anderson, R.P. (2015) spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* 38, 541–545.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24, 38–49.

Guevara, L., Gerstner, B.E., Kass, J.M. & Anderson, R.P. (2018) Toward ecologically realistic predictions of species distributions: A cross-time example from tropical montane cloud forests. *Global Change Biology* 24, 1511–1522.

Kass, J.M., Vilela, B., Aiello-Lammens, M.E., Muscarella, R., Merow, C. & Anderson, R.P. (2018) *Wallace*: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. *Methods in Ecology and Evolution* 9, 1151–1156.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17, 145–151.

Merow, C., Smith, M.J. & Silander, J.A. (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* 36, 1058–1069.

Muscarella, R., Galante, P.J., Soley-Guardia, M., Boria, R.A., Kass, J.M., Uriarte, M. & Anderson, R.P. (2014) ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods in Ecology and Evolution* 5, 1198–1205.

Naimi, B., Hamm, N.A., Groen, T.A., Skidmore, A.K. & Toxopeus, A.G. (2014) Where is positional uncertainty a problem for species distribution modelling? *Ecography* 37, 191–203.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Townsend Peterson, A. (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in madagascar. *Journal of Biogeography* 34, 102–117.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M. & Araújo, M.B. (2011) *Ecological niches and geographic distributions (MPB49)*, vol. 56. Princeton University Press.

Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E. & Blair, M.E. (2017) Opening the black box: an open-source release of Maxent. *Ecography* 40, 887–893.

Phillips, S.J. & Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* 31, 161–175.

Radosavljevic, A. & Anderson, R.P. (2014) Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography* 41, 629–643.

Shcheglovitova, M. & Anderson, R.P. (2013) Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. *Ecological Modelling* 269, 9–17.

Veloz, S.D. (2009) Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography* 36, 2290–2299.